BayeShield: Conversational Anti-phishing User Interface

Peter Likarish, Don Dunbar, Juan Pablo Hourcade, Eunjin Jung
The University of Iowa
Dept. of Computer Science
15 MLH Iowa City, IA 52242
plikaris@cs.uiowa.edu, donald-dunbar@uiowa.edu, hourcade@cs.uiowa.edu, ejjung@cs.uiowa.edu

1. INTRODUCTION

Phishing is an attack in which users are fooled into entering personal information into a "spoof" website instead of the intended legitimate website. Traditional anti-phishing tools, such as the default tools in Internet Explorer 6+ and Mozilla Firefox 2.0/3.0, rely primarily on blacklists, lists of URLs that have been observed hosting phishing attacks. Blacklists provide no protection from attacks that are not already flagged as phishing. The number of such missed attacks is considerable [5]. Researchers have proposed supplementing blacklists with Information Retrieval (IR)-based tools [10]. However, an IR-based approach may generate false positives; legitimate websites incorrectly flagged as phishing. False positives may undermine user trust in a tool and pose questions of legal liability.

We are investigating whether or not a conversational user interface (UI) can help users determine if a website is phishing or legitimate. Our anti-phishing UI, BayeShield, can serve as the front-end to an IR-based tool that identifies phishing attacks with high probability but may produce a small number of false positives. First, we discuss related work and follow this with our system design. We then conclude by detailing the highlights of a formative user study we have conducted and outline future work.

2. RELATED WORK

Several studies have explored why people are vulnerable to phishing attacks and which UI indicators successfully warn people when they are at risk of an attack. Studies by Wu et al. [7] and Dhamija et al. [3] found users do not notice passive indicators. More recently, Egelman et al. compared the default anti-phishing tools in IE7 and FF2.0 and found 1) active indicators prevented phishing significantly better and 2) their participants found the FF2.0 warning more understandable [4]. Various anti-phishing UIs have been proposed, including Passpet by Yee et al. [9], Dynamic Security Skins by Dhamija and Tyger [2], and Web Wallet by Wu et al [8]. iTrustPage by Ronda et al. is perhaps the most similar approach to ours [6]. iTrustPage leverages user expertise in visual tasks whereas we ask the user questions that are difficult or impossible for a software program to determine.

3. SYSTEM DESIGN

Although IR-based techniques are highly successful at identifying phishing websites [10], it is difficult to eliminate false positives without contextual knowledge of how users reached the website and what they intend to do at the site. BayeShield's goal is to partner with users to make a decision about the risk of entering information on a suspicious website, and in the process, educate users about the risk factors associated with phishing websites.

We leverage the fact that users are cognizant of the actions that lead them to a suspicious website by asking them simple questions. This approach is justified by research from Brustoloni and Villamarin-Salomón, who discovered that polymorphic, context sensitive guidance (CSG) reduces the risk users are willing to take in a security setting [1]. In our case, we use CSG to inform users they are at risk of falling for a phishing attack and to leverage their contextual knowledge to correctly judge the potential site in question. Future research may compare displaying options in a polymorphic fashion versus the current static order.

When the anti-phishing tool's detection algorithm determines a website is possibly phishing, an overlay and pop-up (similar to FF2.0's anti-phishing UI) prevents the user from proceeding. The user is then asked to answer a series of questions from the "BayeShield Analyzer" to determine if it is safe to proceed.

Table 1.One example of a series of Analyzer questions.

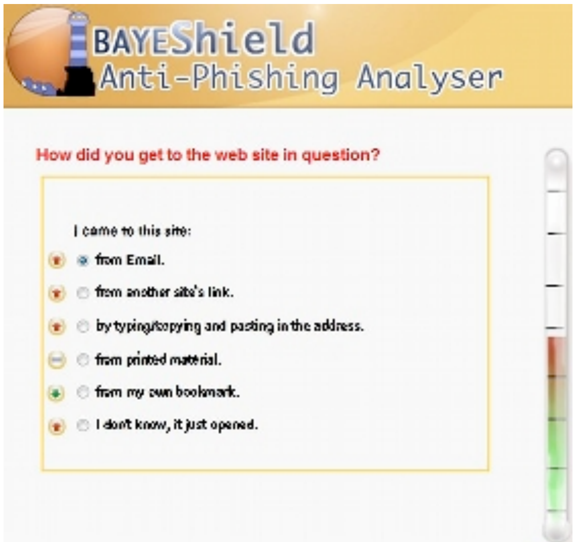
Questions	User's Response
How did you get to the site?	From email
Do you recognize the company/person?	Yes
This email was:	Unexpected
Did the email convey a sense of urgency?	Yes

If the user agrees, they are presented with a wizard-style pop-up designed to be professional and calming. This Analyzer explains they may have arrived at a dangerous website and will be asked a series of questions. In all cases, the Analyzer speaks the user's language, using clear sentence construction with definitions available for technical terms.

The questions walk users through a decision tree modeled on how an expert would decide whether a website is dangerous or not. Due to a lack of space, we cannot include the entire tree but summarize a path through the decision tree in Table 1. Fig. 1 displays one of the questions asked. In response to their answers, a meter at the right will raise or lower. This visually conveys the risk associated with their answers: the higher the bar, the more dangerous it is for them to enter their personal information. In Fig. 1, notice we provide indicators next to each answer signifying whether selecting that answer will increase or decrease the risk.

The first question asks what information the website requests. Users select from four categories: identity, personal info, financial info or account info. Each category has an icon to visually convey the category type. Next, users are asked how they arrived at the

Symposium On Usable Privacy and Security (SOUPS) 2009, July 15-17, 2009, Mountain View, CA, USA.



How did you get to the web site in question?

I came to this site:

● from Email.

○ from another site's link.

○ by typing/copying and pasting in the address.

○ from printed material.

○ from my own bookmark.

○ I don't know, it just opened.

Figure 1 The Analyzer asks how the user reached the site. Their answer alters the meter's height, visually conveying the threat level.

site in question (Fig. 1). At this point, the questions diverge depending on their answers. After answering questions, a summary screen informs them if it is likely to be safe to enter their information (or alternatively, unsafe to do so).

4. USER STUDY

In order to evaluate the usefulness of our UI we conducted a formative study with 20 participants.

4.1 Participants

Potential participants were directed to an online survey. All 20 who completed the survey were selected to participate in the in-lab session. Fourteen were female, six male whose age ranging from 19 to 46 (27.6 median). All participants completed some college and 8, some graduate work. The plurality used FF2.0, followed by IE. They used computers an average of 5.5 hours a day (1.5 hours online). Each participant completed one 45-minute session in an on-campus lab between Sep. 1st and 9th, 2008.

4.2 Design

We asked participants to complete a series of tasks designed to evaluate whether they could distinguish between phishing attacks and false positives after answering questions from BayeShield's Analyzer. The tasks were as follows:

- Email: The participant clicks on a link in an email warning them their account will be disabled in 24 hours if they do not log-in. (phishing task)
- Copy/Paste: The participant pastes a URL containing a misspelling (www.bank0famerica.com). (phishing task)
- Brochure: The participant picks up a brochure after visiting a state park. The brochure contains a URL, they type it to donate to the park. (false positive task)
- Bookmark: The participant selects a bookmark to their stockbroker. After using the Analyzer, BayeShield misinforms them, telling them it is not safe to proceed due to the amount of information requested by the site. (false positive task)

5. RESULTS

We highlight the results of this study. Encouragingly, 19 of 20 participants used the Analyzer correctly on the Email task and *no user entered information after using the Analyzer on the email task*. 75% of users noticed the homograph attack in the copy/paste task after using BayeShield. For the Brochure task, we predicted most participants would select "from printed material" when asked how they arrived at the site but many participants selected "typing or copy/paste." As a result, BayeShield correctly identified the site as "safe" only 65% of the time. Despite this, 16 participants correctly identified the site as safe. In the bookmark task, BayeShield incorrectly informed the users that the site was unsafe when it was actually a site they had bookmarked and the warning was a false positive. Still, 65% of participants correctly identified this as a false positive. We are working on improving the wording of the questions to improve false positive recognition.

6. CONCLUSION

We believe a conversational UI can help users distinguish between risky and safe behaviors online by teaching them the risks associated with their actions. This may allow advanced IR-based anti-phishing tools that may produce occasional false positives to be used in real-life applications. We intend to conduct a longitudinal study to determine if the conversational approach has an educational effect and to further investigate the encouraging findings of our initial study.

7. REFERENCES

[1] Brustoloni, J C, Villamarin-Salomón. Improving Security Decisions with Polymorphic and Audited Dialogs. SOUPS 2007, ACM Press (2007), 77-87

[2] Dhamija, R, Tygar J D. The Battle Against Phishing: Dynamic Security Skins. Proc. SOUPS 2005, ACM Press (2005), 77-88

[3] Dhamija, R, Tygar, J D, Hearst, M. Why Phishing Works. Proc. CHI 2006, ACM Press (2006), 581-590.

[4] Egelman, S., Cranor, L.F., Hong, J. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. Proc. CHI 2008, ACM Press (2008), 1065-1074.

[5] Ludl, C., McAllister, S., Kirda, E., Kruegel, C. On the Effectiveness of Techniques to Detect Phishing Sites. Detection of Intrusions and Malware, Spring (2007), 20-39

[6] Ronda, T., Saroiu, S., and Wolman, A. iTrustPage: A User-Assisted Anti-Phishing Tool. Proc. EuroSys'08.

[7] Wu, M., Miller, R., Garfinkel, S. Do Security Toolbars Actually Prevent Phishing Attacks? Proc. CHI 2006, ACM Press (2006), 601-610.

[8] Wu, M., Miller R., Little, G. Web Wallet: Preventing Phishing Attacks by Revealing User Intentions. Proc. SOUPS 2006, ACM Press (2006), 102-113.

[9] Yee, K.P., Sitaker K. Passpet: Convenient Password Management and Phishing Protection. Proc. SOUPS 2006, ACM Press (2006), 32-43.

[10] Zhang, Y., Hong, J., Cranor, L. Cantina: A Content-based Approach to Detecting Phishing Websites. Proc. WWW07, (2007), 639-64